

Lots of Data – a perspective from an academic publisher

Presented by: Sweitze Roffel, sr. Publisher, Elsevier

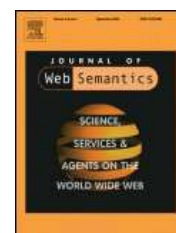
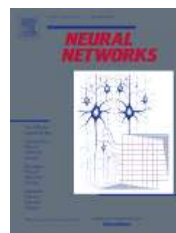
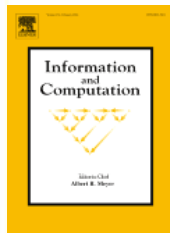
Venue: ISO/IEC Study Group on Big Data, CWI, Amsterdam, the Netherlands , May 14, 2014



Introduction



- Sweitze Roffel
- With Elsevier as a Publisher since 2004
- Currently responsible for a number of publications ranging from Artificial Intelligence to Theoretical Computer Science
 - Proud sponsor of the Semantic Web Challenge including Big Data Track
 - <http://challenge.semanticweb.org/> -



Outline



- Some history
- To the future
- Linking out of a publishing house
- Linking into a publishing house
- Linking to deeper knowledge
- Linking infrastructure
- Linking all around

Some History...



Original House of Elzevier ca. 1580

Based in Leyden as printer to the University

Publisher of early landmark works in science



Including

Galileo Galilei's
Discorsi e Dimostrazioni Matematiche, Intorno a Due Nuove Scienze
(Two New Sciences).
1638

We're not in Leiden any more

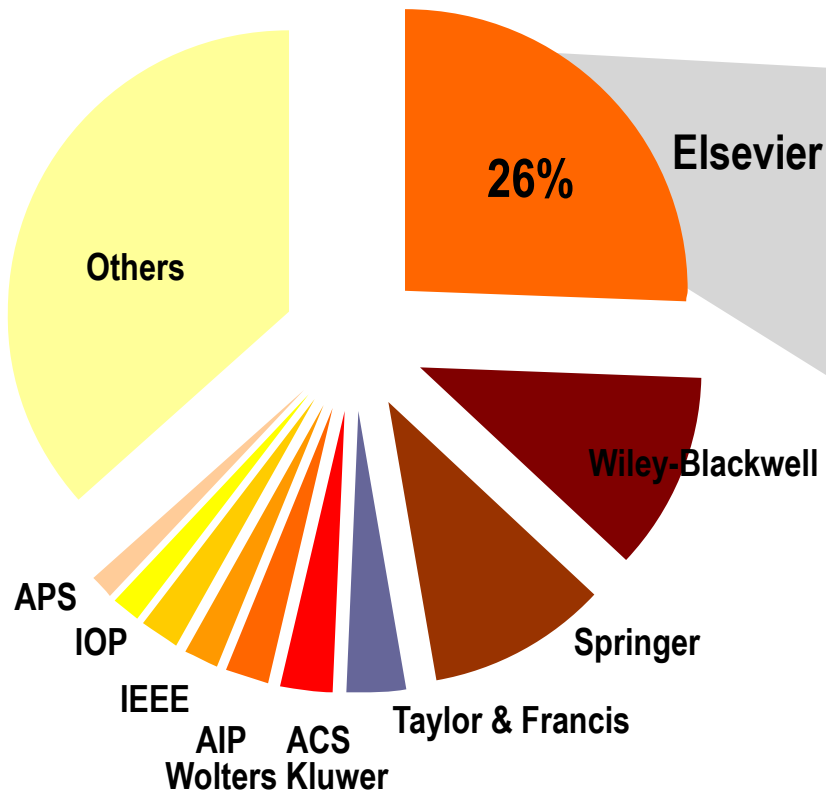


From only a few in 1880, modern Elsevier grew to over 7000 employees serving our customers worldwide today

Elsevier's journal programme today

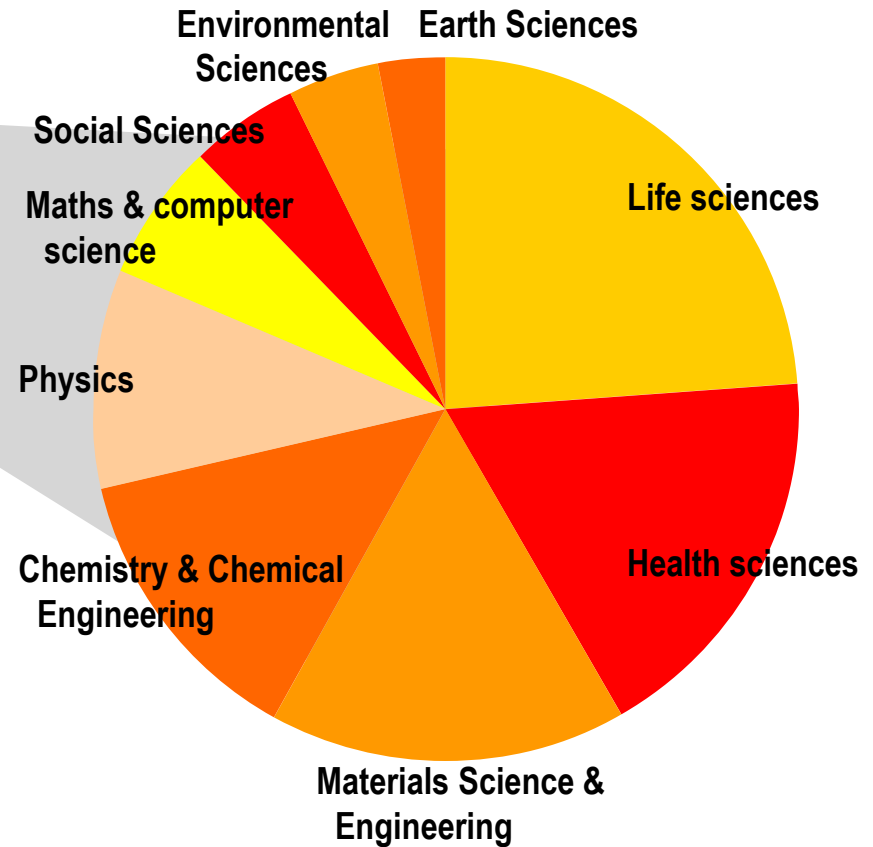


Share of Journal Articles Published



Over 1 ½ million English language research articles published globally each year

Our Scientific Disciplines



About 1000 English language research articles published with Elsevier today
(i.e. 1000 per day)


big picture: your basic options to disseminate new knowledge into the world



1. Keep it secret, and build a service or product using new knowledge
 - NOT Letting others know about it directly
 - Others need to guess / reverse engineer knowledge based on product or service
 - Not adding to 'prior art' (others can still patent or publish this knowledge)
2. Apply for patent
 - Let others know about it directly – establish priority
 - Commercial protection of your new knowledge (IP, licensing etc.)
 - Adds to 'prior art' (others can no longer patent or publish this knowledge)
3. Publish in the scientific literature
 - Let others know about it directly - establish priority
 - No commercial protection of your new knowledge
 - Adds to 'prior art' (others can no longer patent or publish this knowledge)

Example of mixed IP approach





Hub | **ScienceDirect** | Scopus | Applications

Register | Login | Go to SciVal Suite



Home | Browse | Search | My settings | My alerts

Back to results | Browse | 1 of 1 | Next » | Export citation | PDF (1457 K) | More options...

**Computer Networks and ISDN Systems**
Volume 30, Issues 1–7, April 1998, Pages 107–117
Proceedings of the Seventh International World Wide Web Conference



The anatomy of a large-scale hypertextual Web search engine [☆]

Sergey Brin  , Lawrence Page 

Computer Science Department, Stanford University, Stanford, CA 94305, USA

Available online 17 June 1999.

[http://dx.doi.org/10.1016/S0169-7552\(98\)00110-X](http://dx.doi.org/10.1016/S0169-7552(98)00110-X); How to Cite or Link Using DOI

Cited in by Scopus (1946)

 Permissions & Reprints

Abstract

In this paper, we present **Google**, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. **Google** is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at <http://google.stanford.edu/>


To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of Web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the Web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and Web proliferation, creating a Web search engine today is very different from three years ago. This paper provides an in-depth description of our large-scale Web search engine — the first such detailed public description we know of to date.

Apart from the problems of scaling traditional search techniques to data of this magnitude, there are new technical challenges involved with using the additional information present in hypertext to produce better search results. This paper addresses this question of how to build a practical large-scale system which can exploit the additional information present in hypertext. Also we look at the problem of how to effectively deal with uncontrolled hypertext collections where anyone can publish anything they want.

Related articles

- Web dynamics and their ramifications for the...
Computer Networks
- Web search engine multimedia functionality
Information Processing & Management
- Web search enhancement by mining user action...
Information Sciences

[View more related articles](#)


 **Table Download**

Find HTML data tables from the current article to download.

[Find Tables](#) 

[About Table Download](#)

 **Share**

[citeulike](#)  [Like](#) [Tweet](#)

[Add apps](#) | [Help](#)

Related reference work articles
e.g. encyclopedias

- Web Searching
Encyclopedia of Language & Linguistics (Seco...
- Search Engines

Google founders published one bit (PageRank)

and kept quiet about the other bit (AdWords)

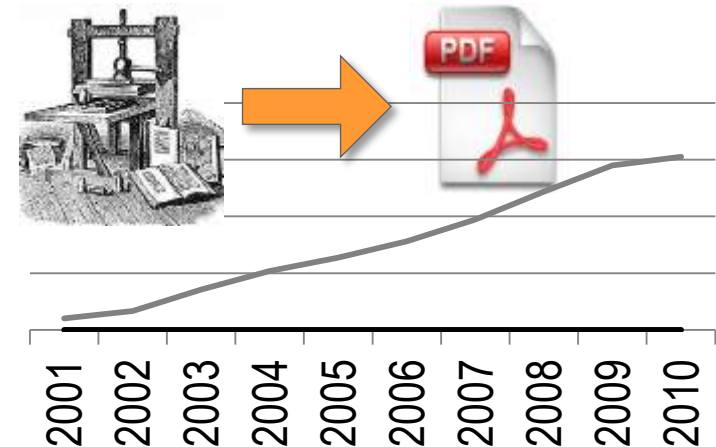
Generating some IP revenue for Stanford

Some more “recent” changes in publishing

In the 1980's arrival of desktop publishing, SGML, etc. “invisible” back end made electronic, front end for LaTeX users -but still dissemination through print .

In the late 1990's Internet enabled move from Print dissemination to Online...

- New functionality: linking, searching
- Improved distribution and access
- Separation of content and functionality
- Branded platforms for e-publishing (Science Direct, High Wire Press, Springerlink)
- Re-Digitized 350+ years of historical printed articles back to vol 1 /issue 1 (back files)
- New disruptive sales channels (amazon for books, etc.)
- New business models: Big deals, Open Access, Sponsored Articles, Pay per View, Collections,
- Manuscript submission and peer review also moved online



However, the format of the traditional article had not really adapted

Scientific publishing perceived as squeezing complex, multi-dimensional research onto the traditional format: *“a rectangular area with ink on a piece of paper”*

Elsevier as a partner and as a platform



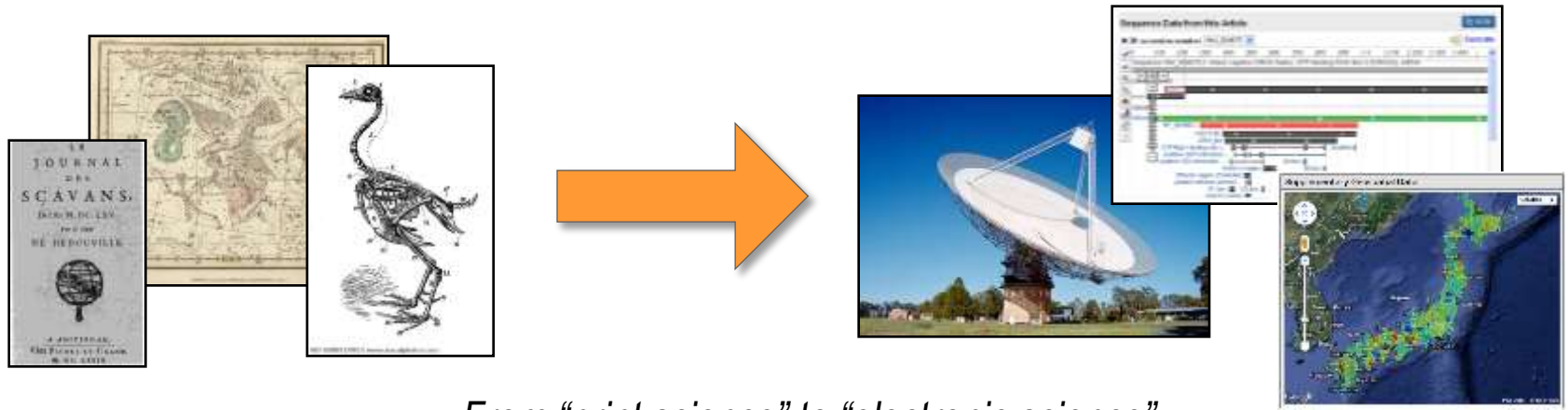
- The next online phase will be about adding intelligence to the process
 - And that means digitalizing 'knowledge artifacts' better – and make them flow.
- We fully realize we can't possibly build all of the future solutions ourselves
 - That means partnering with
 - editors, authors, reviewers, end users, librarians,
 - governments, funding bodies, deans, provosts, R&D managers, societies, other publishers
 - developers, tech companies, standards bodies ,
- So we need to be able to collaborate - STRUCTURALLY
 - Technically, operationally and organizationally
 - In a heterogeneous environment
 - With respect for everyone's property and rights
 - Under global budget constraints
- How?
 - Fact based approach to the future:
 - Investigate, test , learn, pilot, before implementing and scaling
- To deliver solutions that work well
 - are overall cost effective
 - are used and liked
 - and are financially, technically and academically sustainable

Outline



- Some history
- To the future
- Linking out of a publishing house
- Linking into a publishing house
- Linking to deeper knowledge
- Linking infrastructure
- Linking all around

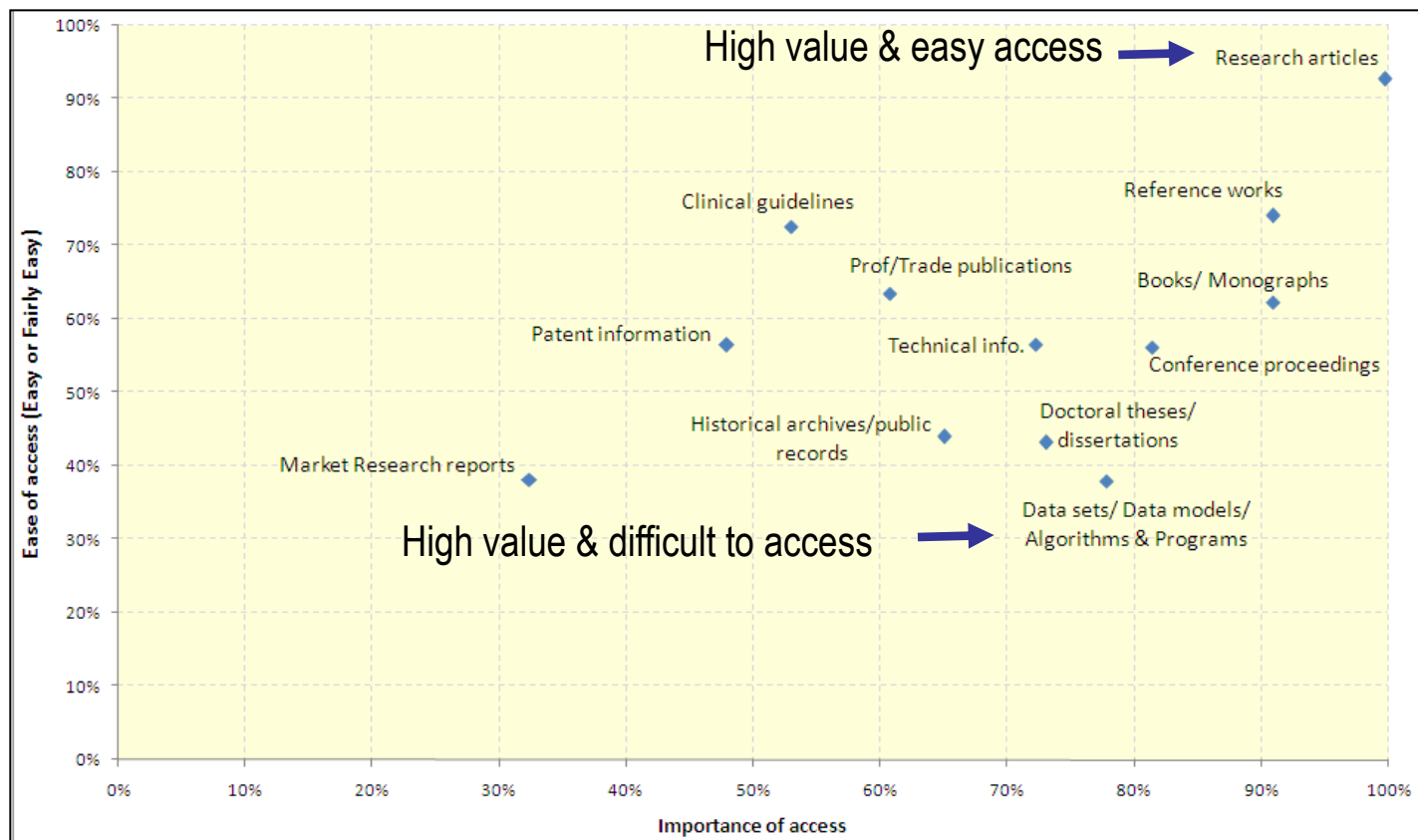
From “print science” to “electronic science” ...



From “print science” to “electronic science”

- Researchers use digital tools to gather, analyze, share data
- Research output includes text, figures, multimedia, data, code, ...
- New tools to share, visualize, and interact with information

What do researcher's themselves say?



Source: independent global study commissioned by Elsevier.
4,109 in depth researcher respondents (almost 7% of apx 60.000)

Interlinking research Articles and research Data adds value both ways



- Increase visibility, discoverability, and usage
- Provide context, avoid misinterpretation and incorrect usage
- Ensure long-term availability of useful content and context
- Coordinate submission process / deposit mechanism

85% of researchers believe it is useful to link underlying digital research data to the formal literature (PARSE.Insight)

Three components of the Article of the Future concept:

- Presentation: Offering an optimal online browsing and reading experience
- Content: Support authors to share digital research output - data, computer code, multimedia files, etc.
- Context: Connecting the online article to trustworthy scientific resources on the web, such as data repositories



Article of the Future a three-pane format



Fun with F1 - Article of the Future - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Article of the Future

www.articleofthefuture.com/10022314098001756-2/

PDF (60 pages) E-mail Export More Display mode

Copyright © 2012 Elsevier B.V. All rights reserved.

View other article Take our survey

Outline Show/hide Outline Article log

Research highlights

Abstract

1. Introduction

2. The abelian part of the BC-system and its endomorphisms

2.1 Group-theoretic description

2.2 The endomorphisms ρ_n from algebraic geometry

3. \mathbb{F}_1^* and the abelian part of the BC-system

3.1 Arithmetical varieties over \mathbb{F}_1

3.2 The varieties $\mu^{(k)}$

4. The integral BC-endomorphisms

4.1 C^* -algebraic description of the BC-system

4.2 The BC-algebra over \mathbb{Q}

4.3 The maps ρ_n

4.4 The BC-algebra over \mathbb{Z}

4.5 Relation with the integral Hecke algebra

5. The endomorphisms and algebras in characteristic p

5.1 The endomorphisms in characteristic p

5.2 The BC-algebra in characteristic p

5.3 The effect of reduction

5.4 Endomorphisms in the unramified case

Video is also available at: http://www.youtube.com/watch?v=ac_djpcv710

Video: Alain Connes explains his paper

1. Introduction

Starting with seminal observations of J. Tits on the classification of simple finite groups (cf. [10]), the a priori vague idea that a suitable analogue of the geometry over the finite fields \mathbb{F}_q should make sense in the limit case $q = 1$ has been taking more and more substance and has given rise to a number of different approaches (cf. [9], [10], [11], [12], [15], [16] and [19]). So far, the relation between these constructions and the Riemann zeta function has remained elusive, in spite of the hope of being able to consider the tensor product $\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}$ as a non-trivial analogue of the product of a curve by itself (see [12]).

It is known that the quantum statistical mechanical system of [1] (which we refer to as the BC-system) gives, after passing to the dual system, a spectral realization of the zeros of the Riemann zeta function, as well as a trace formula.

The main result of [4] applied to the BC-system is the following theorem.

Theorem 6.3 The structure of the BC-endomorphisms corresponds to the structure of \mathbb{F}_1^* over \mathbb{F}_1 as follows:

a) The abelian part of the BC-endomorphisms over \mathbb{F}_1 corresponds to the inductive system of "adelic" \mathbb{F}_1^* .

b) The endomorphisms σ_n describe the Frobenius correspondence, in the sense that on the algebra $\mathbb{Z}[\mathbb{Q}/\mathbb{Z}] \otimes_{\mathbb{Z}} \mathbb{K}$, for \mathbb{K} a perfect field of characteristic $p > 0$, the endomorphisms σ_n coincide with the Frobenius correspondence described in Remark 6.2.

Left pane: efficient navigation & browsing

Center pane: "Traditional" full-text view, designed for optimal online reading experience

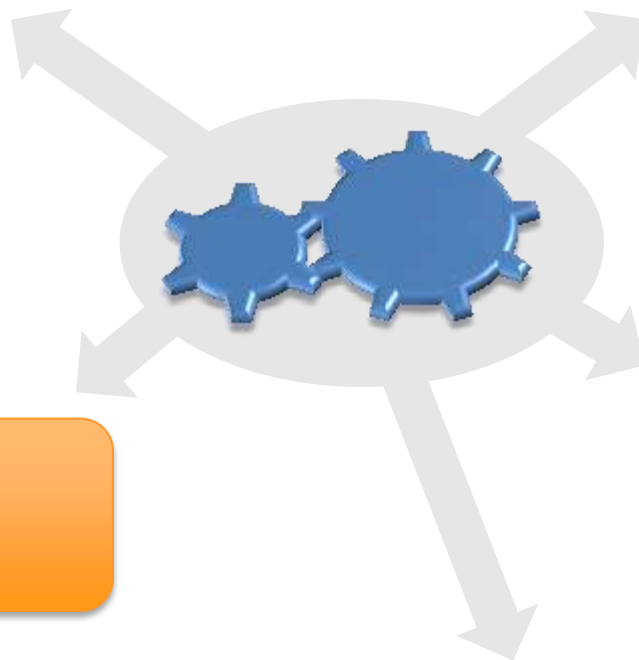
Right pane: Additional content & tools. Shown here: theorem browser

Content Innovations for Supplementary Data



Data viewers built into
ScienceDirect

Support for new kinds of data



Inline Supplementary
Material

Data pilots

The Executable Paper

Outline



- Some history
- To the future
- Linking out of a publishing house
- Linking into a publishing house
- Linking to deeper knowledge
- Linking infrastructure
- Linking all around


Linking out of a publishing house to data



- Much relevant Research Data lives in external subject specific databases
- Repositories of varying nature
 - Many subject areas, (and sub-sub-subject area's)
 - Different scopes & data practices,
 - Structure, Terminology, ontology, ambiguity ,
 - Systems, Content formats, identifiers, data structure
 - Size, Usage
 - Organization, Business models
 - Access ,Legal rights
 - Value (in Science relatively small usage, size, or scope may still be extremely relevant)
 - Etc...(many V's)

Example of Linking Articles to external data repositories



 Download PDF  Export citation  Jump to references  More options...

 **Physics Reports**
Volume 399, Issues 2–3, September 2004, Pages 71–174



Studies of hadronic event structure in e^+e^- annihilation from 30 to 209 GeV with the L3 detector

P. Achard^a, O. Adriani^b, M. Aguilar-Benitez^{c,1}, J. Alcaraz^{c,1}, G. Alemani^d, J. Allaby^e, A. Aloisio^f, M.G. Alvigi^g, H. Anderhub^g, V.P. Andreev^{h,i}, F. Anselmo^j, A. Arefiev^k, T. Azemoon^l, T. Aziz^m, P. Bagnaiaⁿ, A. Bajo^{c,1}, G. Baksay^o, L. Baksay^o, S.V. Baldew^p, S. Banerjee^m, Sw. Banerjee^q, A. Barczyk^{q,r}, R. Barillère^e, P. Bartalini^d, M. Basile^j, N. Batalova^s, R. Battiston^t, A. Bay^d, F. Becattini^b, U. Becker^u, F. Behner^q, L. Bellucci^b, R. Berbeco^l, J. Berdugo^{c,1}, P. Berges^u, B. Bertucci^t, B.L. Betev^q, M. Biasini^t, M. Biglietti^f, A. Biland^q, J.J. Blaising^q, S.C. Blyth^v, G.J. Bobbink^p, A. Böhm^{w,2}, L. Boldizsar^{x,3}, B. Borgiaⁿ, S. Bottai^b, D. Bourilkov^q, M. Bourquin^a, S. Braccini^a,
I.G. Branson^y, F. Brochu^q, I.D. Burner^u, W.I. Burner^t, X.D. Cai^u, M. Canell^u, G. Cara Romeo^j, G.

<http://dx.doi.org/10.1016/j.physrep.2004.07.002>

Bibliographic information

Citing and recommended articles

Applications and tools

Data for this Article

[More information on this application](#)

Data for this article is available at the following data repositories:

 **HepData**
View reaction data from this article at the Durham Reaction Database

Workspace

<http://www.sciencedirect.com/science/article/pii/S0370157304002753>

Interlinking Articles and Data through banners



 Download PDF  Export citation  Jump to references  More options...

 **Physics Reports**
Volume 399, Issues 2–3, September 2004, Pages 71–174



Studies of hadronic event structure in e^+e^- annihilation from 30 to 209 GeV with the L3 detector

P. Achard^a, O. Adriani^b, M. Aguilar-Benitez^{c,1}, J. Alcaraz^{c,1}, G. Alemani^d, J. Allaby^e, A. Aloisio^f, M.G. Alvigi^g, H. Anderhub^g, V.P. Andreev^{h,i}, F. Anselmo^j, A. Arefiev^k, T. Azemoon^l, T. Aziz^m, P. Bagnaiaⁿ, A. Bajo^{c,1}, G. Baksay^o, L. Baksay^o, S.V. Baldew^p, S. Banerjee^m, Sw. Banerjee^q, A. Barczyk^{q,r}, R. Barillère^e, P. Bartalini^d, M. Basile^j, N. Batalova^s, R. Battiston^t, A. Bay^d, F. Becattini^b, U. Becker^u, F. Behner^q, L. Bellucci^b, R. Berbeco^l, J. Berdugo^{c,1}, P. Berges^u, B. Bertucci^t, B.L. Betev^q, M. Biasini^t, M. Biglietti^f, A. Biland^q, J.J. Blaising^q, S.C. Blyth^v, G.J. Bobbink^p, A. Böhm^{w,2}, L. Boldizsar^{x,3}, B. Borgiaⁿ, S. Bottai^b, D. Bourilkov^q, M. Bourquin^a, S. Braccini^a,
I.G. Branson^y, F. Brochu^q, I.D. Burner^u, W.I. Burner^t, X.D. Cai^u, M. Canell^u, G. Cara Romeo^j, G.

<http://dx.doi.org/10.1016/j.physrep.2004.07.002>

Bibliographic information

Citing and recommended articles

Applications and tools

Data for this Article

[More information on this application](#)

Data for this article is available at the

 **HepData**
View reaction data from this article
at the Durham Reaction Database

Workspace

<http://www.sciencedirect.com/science/article/pii/S0370157304002753>

Interlinking Articles and Data through banners



The Durham HepData Project



[REACTION DATABASE](#) • [DATA REVIEWS](#) • [PARTON DISTRIBUTION FUNCTION SERVER](#) • [OTHER HEP RESOURCES](#)

Reaction Database Full Record Display

View [short record](#) or as: [plain text](#), [AIDA](#), [PyROOT](#), [YODA](#), [ROOT](#), [mpl](#) or [ScaVis](#)

ACHARD 2004 — Studies of hadronic event structure in $e^+ e^-$ annihilation from 30-GeV to 209-GeV with the L3 detector

Experiment: [CERN-LEP-L3 \(L3\)](#)

Published in [PRept. 399,71](#) (DOI:10.1016/j.physrep.2004.07.002)

Preprinted as [CERN-PH-EP/2004-024](#)

Record in: [INSPIRE](#)

CERN-LEP. Comprehensive study of hadronic event shapes and distributions in $E^+ E^-$ interactions from collision energies from 91 to 209 GeV. These data update and supersede many of the L3 results published previously.. This section contains the 2,3,4 and 5 jet fractions for the JADE, Durham(KT) and Cambridge algorithms as a function of their respective jet resolution parameters (YCUT). This section contains the distributions of the event shape variables THRUST, Heavy Jet

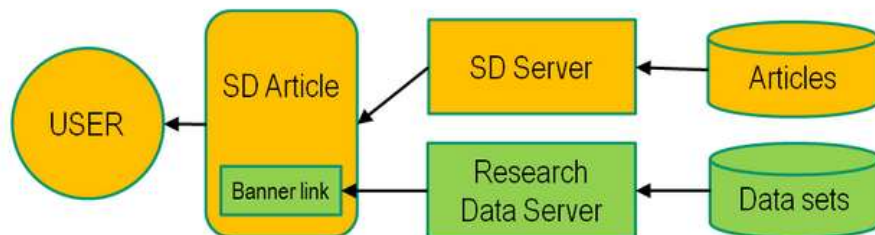
Interlinking Articles and Data through banners



Enabling one-click access to relevant primary data



- Banners linking out to data repositories
- Landing page collects data that is directly relevant for the article
- Enable reproducibility of research, and re-use of data
- For selected data repositories across domains



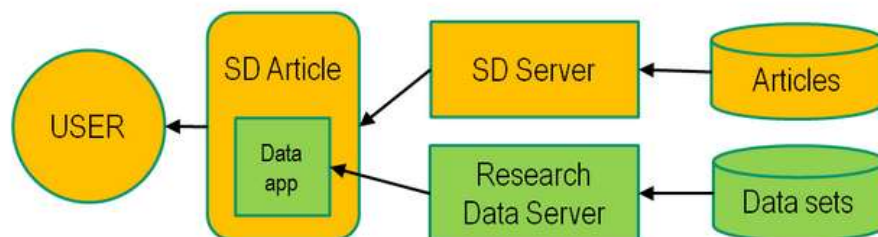
See <http://www.elsevier.com/databaselinking>

Apps serving data integration and visualization tools bring Articles and their data even closer



Integrating (meta)data into the article page view

- Supplementary data at PANGAEA
- Bidirectional links between PANGAEA <> ScienceDirect
- Data visualized next to the article



See <http://www.elsevier.com/databaselinking>

Outline



- Some history
- To the future
- Linking out of a publishing house
- **Linking into a publishing house**
- Linking to deeper knowledge
- Linking infrastructure
- Linking all around

Article of the Future Framework does not reinvent the wheel

Bringing 3rd party functionality into the article to link the data

Turning a static image from the article into an interactive one:



- Present research findings in an more valuable, interactive way
- Help readers find and understand data in the context of the article
- Download data for validation & re-use of data

Data-integration brings Articles and Data even closer



Identification of a New Motif in Family B DNA Polymerases by Mutational Analyses of the Bacteriophage T4 DNA Polymerase

Vincent Li¹, Matthew Hogg², Linda J. Reha-Krantz¹,  

¹ Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2E9

² Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, VT 05405, USA

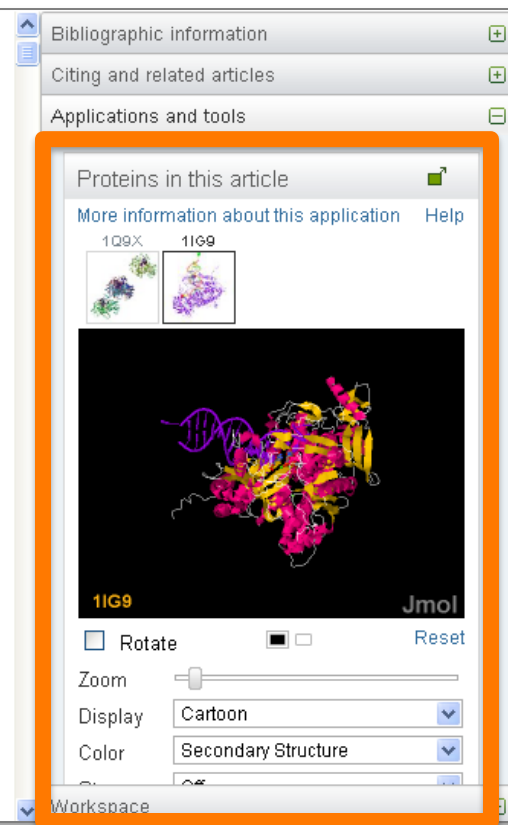
<http://dx.doi.org/10.1016/j.jmb.2010.05.030>, [How to Cite or Link Using DOI](#)

 [Permissions & Reprints](#)

Abstract

- Explore protein structures relevant to the article – zoom, rotate, etc.
- Structure and other protein data integrated from Protein Data Bank
- Author-tagged accession numbers

motif that
been the
of the
that form
lethality
w-dGTP
enerally
that can
ino acid
tivity to
eplicate
that the
revealed




The screenshot displays a web interface for exploring protein structures. On the left, a sidebar contains a list of links: 'Bibliographic information', 'Citing and related articles', and 'Applications and tools'. The main content area is titled 'Proteins in this article' and includes a link for 'More information about this application' and a 'Help' button. Below this, two small thumbnail images of protein structures are shown, labeled '1Q9X' and '1IG9'. The central focus is a large 3D molecular model of a protein structure, labeled '1IG9' in the bottom left corner and 'Jmol' in the bottom right corner. The model is rendered in a cartoon style with a color scheme of pink, yellow, and blue. Below the model, there are interactive controls: a 'Rotate' button, a 'Zoom' slider, a 'Display' dropdown menu set to 'Cartoon', and a 'Color' dropdown menu set to 'Secondary Structure'. A 'Reset' button is also present.

See <http://www.elsevier.com/databaselinking>

Article of the Future partnering to deliver Interactive MATLAB viewer



Making plots more valuable for research



Computer Methods in Applied Mechanics and Engineering
Volumes 245–246, 15 October 2012, Pages 75–89

Integrated layout design of multi-component systems using XFEM and analytical sensitivity analysis

J. Zhang, W.H. Zhang, J.H. Zhu, L. Xia

Engineering Simulation and Aerospace Computing (ESAC), School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China

<http://dx.doi.org/10.1016/j.cma.2012.06.022>, How to Cite or Link Using DOI

[Permissions & Reprints](#)

Abstract

This study presents the integrated layout optimization of multi-component systems using a fixed mesh. The optimization formulation is established under the framework of the extended finite element method (XFEM). The level set method is used to represent components and is combined with the XFEM to describe material discontinuities across elements. Sensitivity analysis is proposed with respect to geometric variables of components and pseudo-densities of the basic structure. An analytical shape sensitivity analysis method

Bibliographic information

Citing and related articles

Applications and tools

Supplementary MATLAB figures

[More information on this application](#) [Disclaimer](#)



Hover the mouse over the image to get access to

[Download figure](#)

Workspace

- Explore figures interactively – zoom, rotate, etc.
- Download underlying data for validation & re-use

Outline



- Some history
- To the future
- Linking out of a publishing house
- Linking into a publishing house
- **Linking to deeper knowledge**
- Linking infrastructure
- Linking all around

Inline Supplementary Data : the case of a simple table



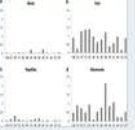
Presenting Supplementary Material at the relevant location

☒ Show thumbnails in outline

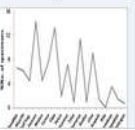
Table 1



4. Results

4.1. Taxonomy




4.2. Taphonomy




 Download PDF  Export citation [Jump to references](#) [More options...](#)

reptiles (Squamata, mainly snakes and lizards), bony fish (Osteichthyes) and birds. The data in Table 1 shows that numbers of specimens are generally low and widely distributed among the samples indicating dispersion rather than concentration of the remains. Using Green's coefficient of dispersion (see Krebs, 1988; Formula (1.95)), the results for the 19 samples are: $G = 1.08$, $G^2 = 1.17$, $G^2/G = 1.08$, indicating a distribution rather than aggregated or uniform distribution of the remains among the samples.


 **Inline Supplementary Table S1**


The average number of specimens per liter of sediment (i.e., density) across the 19 samples is 1.7 with a standard deviation of 0.9 (see densities in Table 1). High density of 3.6 specimens/l was recorded in Locus 10/Q/88 which is an occupational accumulation overlying a stone floor and low density of <1 specimens/l can be seen in samples from Loci 10/Q/112, 10/Q/46 and 10/Q/68 from various context types. Fig. 3 presents a distribution of densities across the 14 loci for each of the microvertebrate classes. This shows low densities of reptile and bird remains (<0.5 specimens/l) and comparatively high densities of mammal and fish remains. Fish remains show the greatest average density (0.8 specimens/l; $sd = 0.4$). Mammal remains show somewhat lower densities and are also less evenly distributed among the loci than the fish remains (average = 0.7 specimens/l; $sd = 0.5$).


a **Birds** **b** **Fish**




Search ScienceDirect


<http://dx.doi.org/10.1016/j.jas.2012.07.001> 


 Get rights and content


Bibliographic information 


Citing and recommended articles 


Recommended articles

Design choices in imaging speech compr...
2012, NeuroImage
 Show more information

Range of bone modifications by human c...
2013, Journal of Archaeological Science
 Show more information

Political ecology of exurban "lifestyle" lan...
2008, Urban Forestry & Urban Greening
 Show more information

Applications and tools 

Workspace 

Presenting Supplementary Material at the relevant location

Download PDF Export citation Jump to references More options...

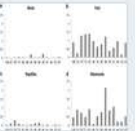
Search ScienceDirect Search

☒ Show thumbnails in outline

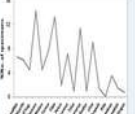
Table 1

4. Results

4.1. Taxonomy



4.2. Taphonomy



reptiles (Squamata, mainly snakes and lizards), bony fish (Osteichthyes) and birds. The data in Table 1 shows that numbers of specimens are generally low and widely distributed among the samples indicating dispersion rather than concentration of the remains. Using Green's coefficient of dispersion (see Krebs, 1989: Equation (4.25)) shows values of 0.09 in Level Q-4 and 0.02 in Level Q-5 which indicate random rather than aggregated or uniform distribution of the remains among the samples.

Inline Supplementary Table S1

Table S1. Counts of specimens by taxonomic categories.

Locus	Basket/level	Bird	Fish
106	3	1	27
112	1	—	5
157	1	—	49
	2	2	80
33	15	—	7
46	16	2	3
	17	13	21
55	6	6	19

http://dx.doi.org/10.1016/j.jas.2012.07.001

Get rights and content

Bibliographic information

Citing and recommended articles

Recommended articles

Design choices in imaging speech compr...
2012; NeuroImage

Show more information

- Supplementary material inserted at the place of reference/citation
- Put material into the right context
- Make it easier for readers to find
- Initially in closed text-box, action to open
- Reader can download to ppt including source, or to CSV for reuse

Digitalizing deeper knowledge

Example from computing : machine consumable code



- For machine consumable code separation of form and content does not work.
- In XML/HTML/ PDF code is often a 'picture' to keep lay out intact
- So we collect actual source code from authors and keep this intact through publication process
- Publish as machine consumable code within article
- Currently piloting this new process on Journal of Web Semantics & Information Sciences



**Please:
with your next
JWS article
please also
submit your
source code**

Inline Supplementary Material (ISM) enriches an article by providing ancillary information in the appropriate context within the main article body. This feature applies to the same kind of material that would otherwise be included as (regular) Supplementary Material, but gives authors the opportunity to make this material much better visible and place it in the right context.

ISM can comprise of the following types of media:

- Figures
- Tables
- Computer code

A pilot was launched in May 2012 for figures in the following journals: Journal of Archaeological Science, Precambrian Research, and for tables in NeuroImage, NeuroImage Clinical and Evaluation and Program Planning. A pilot for Inline Supplementary Computer code on 5 journals is due to start in December 2012.

6. Sharing the goodness

TrialX provides multiple ways for third-party websites to access clinical trial information. We have created a RESTful API through which clinical trials information can be obtained in an RDF output. The RDF information allows general health and wellness websites or bloggers to incorporate clinical trial information enriched with semantic metadata. For example, Web resources that have content on diabetes would automatically be able to pull related clinical trial content from TrialX. [Inline Supplementary Computer Code 1](#) illustrates the RDF export of an Asthma clinical trial.



Inline Supplementary Computer Code 1

RDF export of asthma trial information.

```
<owl:Thing rdf:about="http://trialx.com/#913"/>
- <rdf:Description rdf:about="http://trialx.com/#Asthma">
- <rdf:type>
  <owl:Class rdf:about="http://trialx.com/#Title"></owl:Class>
</rdf:type>
```

Mock-up example of Inline Supplementary Computer Code. Note that this material is presented inside an expandable text box.

Digitalizing deeper knowledge: Example from chemistry where molecules are key artefacts



The screenshot shows the ScienceDirect interface for a journal article. The article title is "Synthesis of 5-arylidene-2-amino-4-azolones and evaluation of their anticancer activity". The authors listed are Ivanna Subtelna, Dmytro Adamanyuk, Ewa Szymańska, Katarzyna Kleś-Kononowicz, Borys Zimenkovsky, Olexandr Vasylenko, Andrzej Gzella, and Roman Lesyk. The article is from Volume 18, Issue 14, 15 July 2010, Pages 5090-5102. The abstract describes the synthesis of novel 5-arylidene-2-arylaminothiazol-4(5H)-ones and 2-aryl(benzyl)amino-1H-imidazol-4(5H)-ones. The supplementary content section shows a download link for MOL files (ZIP file containing the MOL files of the most important compounds in this article).

How does it work?

1. Authors store compound structures as MOL files.
2. Authors upload .MOL files as supplementary material through Elsevier Editorial System (may also be at revision stage)
3. Elsevier turns these MOL files into chemical structures, inserts InChi keys, and links to Reaxys
4. Readers can browse through main compounds, download .MOL files, or click through to Reaxys for more information

Live on ScienceDirect

- Developed together with Reaxys
- Currently live for ~30 journals in chemistry

<http://www.elsevier.com/mol>

Links to highly domain specific knowledge base... ...and Reaxys knowledgebase links back to journal article

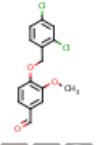
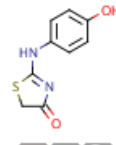
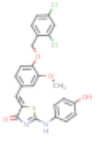


1 reactions out of 1 citations go to Page Page 1 of 1

Filter by:

- Yield
- Record Type
- Reagent/Catalyst
- Solvent
- Reaction Type
- No. of Steps
- Product Availability
- Reactant Availability
- Document Type
- Authors
- Patent Assignee
- Journal Title
- Publication Year

Zoom in Zoom out Hide Sort by Reaxys-Ranking

Yield	Conditions	References
 Create new plan	+	 Create new plan
		 Add this reaction to plan
Rx-ID: 29637768		
95%	With sodium acetate; acetic acid 3 h; Reflux; Knoevenagel condensation;	Subtel'Na, Ivanna; Atamanyuk, Dmytro; Zimenkovsky, Borys; Lesyk, Roman; Szymanska, Ewa; Kiec-Kononowicz, Katarzyna; Vasilenko, Olexandr; Gzella, Andrzej Bioorganic and Medicinal Chemistry, 2010 , vol. 18, # 14 p. 5090 - 5102 Title/Abstract Full Text View citing articles Show Details

Show 9 results per page 1 reactions out of 1 citations go to Page Page 1 of 1

Contact Us | Support | About Reaxys | Terms and Conditions | Privacy Policy | Performance Page
Copyright © 2012 Elsevier Properties SA. All rights reserved. Reaxys® is owned and protected by Elsevier Properties SA and used under license.

- Reaxys allows meaningful navigation through known chemical reactions
- Search on substructures
- Same platform works for everyone – not just Elsevier's Reaxys

Example of content mining - Entity Tagging and Linking





Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids

Volume 1821, Issue 6, June 2012, Pages 884–894



Glycosphingolipid synthesis is essential for MDCK cell differentiation

Lucila G. Pescio^{a, b}, Nicolás O. Favale^{a, b}, María G. Márquez^{b, c}, Norma B. Sterin-Speziale^{a, b},  

^a Cátedra de Biología Celular y Molecular, Departamento de Ciencias Biológicas, Facultad de Farmacia, y Bioquímica, Universidad de Buenos Aires. Ciudad Autónoma de Buenos Aires (C1113AAD), Argentina

^b IQUIFIB-CONICET. Ciudad Autónoma de Buenos Aires (C1113AAD), Argentina

^c Instituto de Investigaciones en Ciencias de la Salud Humana (IICSHUM), Universidad, Nacional de La Rioja. La Rioja. Argentina

Received 5 January 2012. Revised 6 February 2012. Accepted 10 February 2012. Available online 23 February 2012.

<http://dx.doi.org/10.1016/j.bbalip.2012.02.009>, How to Cite or Link Using DOI

Cited by in Scopus (0)

 Permissions & Reprints

Abstract

Glycosphingolipids (GSLs), which are highly concentrated at the apical membrane of polarized epithelial cells, are key components of cell membranes and are involved in a large number of processes. Here, we investigated the ability of hypertonicity (high salt medium) to induce Madin–Darby Canine Kidney (MDCK) cell differentiation and found an increase in GSL synthesis under hypertonic conditions. Then, we investigated the role of GSLs in MDCK cell differentiation induced by hypertonicity by using two approaches. First, cultured cells were depleted of GSLs by exposure to D-*threo*-1-phenyl-2-decanoylamino-3-morpholino-1-propanol (D-PDMP). Second, cells were transfected with an siRNA specific to glucosylceramide synthase, the key enzyme in GSL synthesis. Exposure of cells to both treatments resulted in the impairment of the development of the apical membrane domain and the formation of the primary cilium. Enzymatic inhibitions

Related articles

- Early developmental expression of the gene e...
Biochimica et Biophysica Acta (BBA) - Genera...
- Membrane Dynamics and the Regulation of Epit...
International Review of Cytology
- Tube Morphogenesis: Making and Shaping Biolo...
Cell
- Targeting of recombinant Na⁺/glucose cotrans...
FEBS Letters
- Curcumin inhibits renal cyst formation and e...
European Journal of Pharmacology

▶ View more related articles

Lipid Structures (beta)

7 Instances found

Palmitic Acid

Show Details

Show Occurences - 15

Fumonisin B1

Show Details

Show Occurences - 10

Stearic acid

Show Details

Show Occurences - 8

Cycloserine


Show Details

Show Occurences - 7

Sphingosine

Show Details

Show Occurences - 6

 Cancer Images

Some system issues wrt to sharing



- Not all systems readily talk to each other
- the alphabet is pretty well supported on the web
- So are numbers provided you don't do too much with them
- Data lives in very varied homes
- But what about mathematics? What about computations? How did the author reason with their data?

Further Experimenting and piloting



The screenshot shows the homepage of the 'executable paper grand challenge' website. The header features the Elsevier logo and the title 'executable paper grand challenge' in large, bold letters, with the subtitle 'knowledge enhancement in the computational sciences' below it. A navigation bar includes links for Home, News, About the Challenge, Meet the Winners, Meet the Judges, Abstract Requirements, Finalists, and Rules. The main content area is divided into two columns. The left column, titled 'About the Challenge', contains a paragraph about the importance of data, code, and software in data-intensive research, followed by a list of challenges: Executability, Short and long-term compatibility, Validation, Copyright/licensing, Systems, Size, Provenance, and Other issues. The right column, titled 'Publication', shows a book cover for 'COMPUTATIONAL SCIENCE'. Below this, the 'Organised by' section features the Elsevier logo. The 'Join us on' section includes a Facebook link. The footer contains copyright information and a list of links: Home, About the Challenge, Abstract Requirements, Finalists, News, Intellectual Property Rights, and Contact Us.

executable paper grand challenge
knowledge enhancement in the computational sciences

Home News About the Challenge Meet the Winners Meet the Judges Abstract Requirements Finalists Rules

About the Challenge

Data sets, code, and software are but some of the crucial elements in data intensive research; yet, these elements are noticeably absent when the research is recorded and preserved in perpetuity by way of a scholarly journal article. Further, most researchers do not deposit data related to their research article; and if they do so, it is often deposited on their personal or institutional websites, lacking consistency, reliable dissemination, discoverability, proper association (to the research article), documentation, validation, and preservation. To address all these concerns and to accommodate the every increasing body of data intensive science, considerable adaptations to the existing journal article are fundamental to accommodating the need to disseminate, validate, and archive research data, as well as a method to allow this data, in some way or form, to be validated, citable, tractable, and executable. To achieve this adaptation to scholarly publication, several issues must be addressed; the most vital being:

- Executability**
How to make equations, tables and graphs interactive in such a way that reviewers and readers can check, manipulate and explore the result space? How to make the components of the experiment that generates these elements available to the reader so the experiment can be repeated and manipulated?
- Short and long-term compatibility**
How can we develop a model for executable files that is compatible with the user's operating system and architecture and adaptable to future systems?
- Validation**
How do we validate data and code, and decrease the reviewer's workload? How can validation of this information be made easy for the reviewer?
- Copyright/licensing**
Data of this nature should always be freely available to researchers; how can this principle be encouraged, while maintaining author's patent and intellectual property protection?
- Systems**
How do we convey work done on large-scale computers, which are possibly only available to a small portion of the author/reader community?
- Size**
How do we manage very large file sizes?
- Provenance**
How to support registering and tracking of actions taken on the 'executable paper'?
- Other issues**
How do we tackle risks and liabilities, including viruses and code contamination, plagiarism, or other problems?

Publication

Organised by

Join us on

Copyright © 2011 Elsevier Ltd. All rights reserved. Privacy | Terms & Conditions | Cookies

Home | About the Challenge | Abstract Requirements | Finalists | News | Intellectual Property Rights | Contact Us

The executable paper challenge explores new idea's to digitalize CS research by 'uploading' the whole experiment - aiding reproducibility of research

Pilot investigates aspects of

- Executability
- Compatibility
- Validation
- Licensing
- Systems
- Big Data
- Data *and* Computational Providence

Article of the Future: Executable Papers Pilots



executable paper
grand challenge

Special Issue
Published

PDF (332 K) | Export citation | E-mail article | Highlight keywords on

Article | Figures/Tables (6) | References (15) | Thumbnail

Forest Ecology and Management
Volume 258, Issue 5, 20 August 2009, Pages 722-727

doi:10.1016/j.foreco.2009.05.009 | How to Cite or Link Using DOI
Permissions & Reprints

Carbon concentration variability of 10 Chinese temperate species

Quanzhi Zhang, Chuankuan Wang, Xingchang Wang, Xiankui Quan

College of Forestry, Northeast Forestry University, 26 Hexing Road, Harbin 150040, China

Received 10 March 2009; revised 2 May 2009; Accepted 8 May 2009. Available online 5 June 2009.

Abstract

A mass-based carbon (C) concentration ([C]) of 50% in dry wood is widely accepted as a constant factor of biomass to C stock. However, the [C] varies with tree species, and few data on [C] are available for temperate tree species. In this study, we examined inter- and intra-specific variations of [C] in biomass co-occurring temperate tree species in northeastern China. The species were Korean pine (*Pinus koraiensis* Zucc.), Dahurian larch (*Larix gmelinii* Rupr.), Mongolian oak (*Quercus mongolica* Fisch.), white birch (*Betula papyrifera* Suk.), Amur cork-tree (*Phellodendron amurense* Rupr.), Manchurian walnut (*Juglans mandshurica* Meisn.), Manchurian ash (*Fraxinus mandshurica* Rupr.), aspen (*Populus davidiana* Dode), Mono maple (*Acer mono* Rupr.) and Amur linden (*Tilia amurensis* Rupr.). The mean tissue [C] across the species varied from 47.1% in foliage. The mean stem [C] of the 10 species was $49.9 \pm 1.3\%$ (mean \pm SE). The weighted mean concentration (WMCC) for the species ranked as: Amur cork-tree (55.1%) > Amur linden (53.9%) > Korean pine (53.2%) > Manchurian ash (52.9%) > Manchurian walnut (52.4%) > Mongolian oak (47.6%) > Dahurian Mono maple (46.4%) > white birch (46.1%) > aspen (43.7%). The WMCC of the dominant trees was not correlated to mean annual increment of biomass (MAI), suggesting that planting fast-growing tree species

Collage

Asset 3. Pore center coordinates: row per each pore, columns for x and y coordinates, respectively.
172.0 1.0
93.0 2.0
119.0 2.0
159.0 2.0
311.0 2.0
331.0 2.0
346.0 2.0
490.0 2.0
520.0 2.0
278.0 3.0

Snippet 1 Python 2.6.4

```
Code Output
image_width_in_nm = 10000
image_width_in_px = 10000

input_file_name = './nanostruct/input.in'

area_column_index = 1
circ_column_index = 2
x_coord_column_index = 3
y_coord_column_index = 4
```

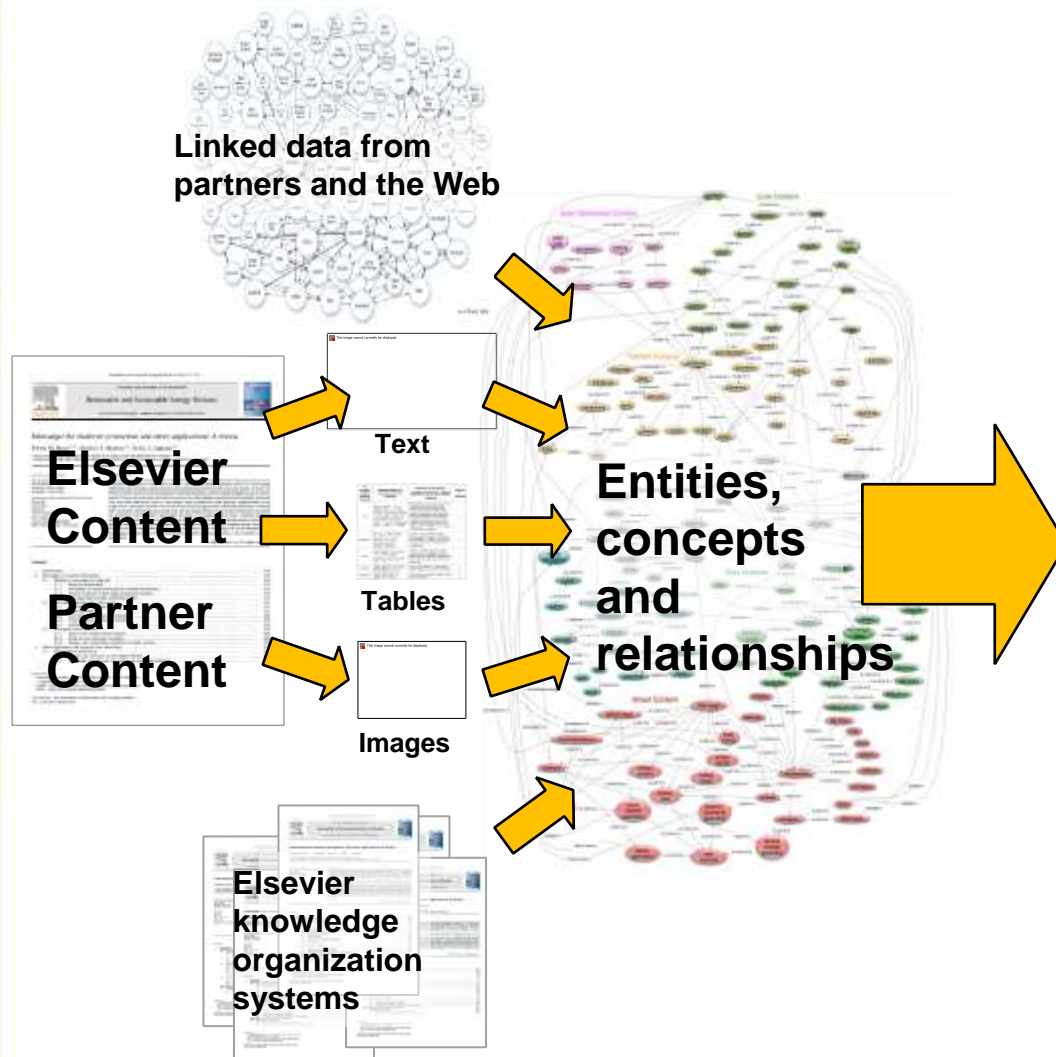
Asset 13. Domain partitioning with respect to tolerance values.

- Opening up the “black box” of computational methods
- Integrate executable components with journal articles
- Pilot Special Issue for Computers & Graphics

Outline



- Some history
- To the future
- Linking out of a publishing house
- Linking into a publishing house
- Linking to deeper knowledge
- **Linking infrastructure**
- Linking all around

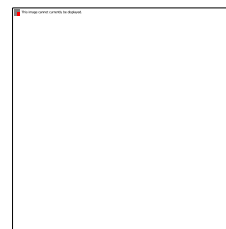


Smart Content Applications



Better discovery through semantic search & navigation

- Faceted search & browse
- Ontology-driven navigation
- Task-specific results
- Personalized/localized results
- Link to evidenced-based content



Better understanding through analysis and visualization

- Question & Answer
- Actionable Content & Alerts
- Tag clouds
- Heatmaps
- Animations



New knowledge through aggregation and synthesis

- Topic pages
- Social network maps
- Geolocation maps
- Data integration and mashups
- Text mining
- Inference and Reasoning

Linked Data Repository (LDR): Elsevier's Warehouse for Smart Content Enhancements



Delirium treatment: An unmet challenge

Title

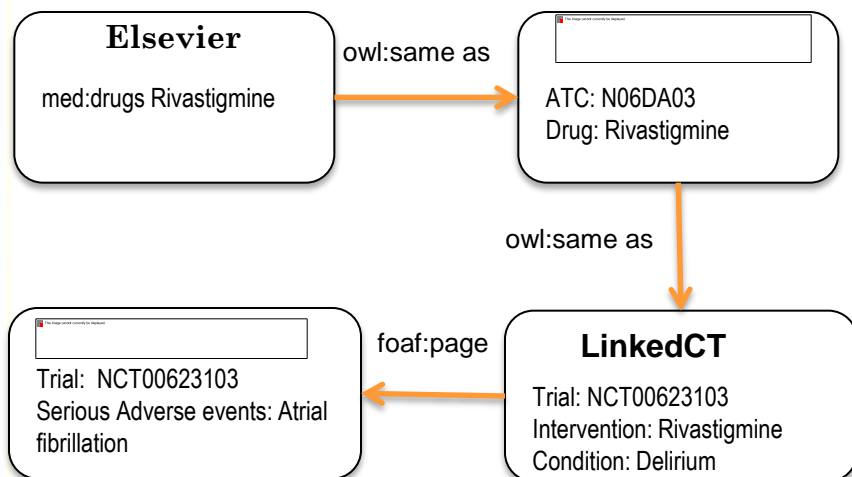
Rivastigmine, a cholinesterase inhibitor, has been used to treat delirium in elderly patients with stroke. 1 A biologically plausible premise is that enhanced cholinergic transmission may counteract the cholinergic deficit in delirium—led to a placebo-controlled, double-blind trial by Maarten van Eijk and colleagues.

Drugs

Disease

Clinical finding

they added rivastigmine or placebo to the treatment of patients in intensive care. The trial was halted at 104 patients by the drug safety and monitoring board (DSMB) because of increased mortality (12/54 in the rivastigmine group, 4/50 in the placebo group; $p=0.07$) and a worse outcome. The rivastigmine group ...



- Knowledgebase of semantic data for deeper insight via exploration, analysis, and visualization
- Large scale integration of related sources of medical and scientific content and data
- Provides high performance service layer APIs for ease of integration into end-user products and other platforms

Outline



- Some history
- To the future
- Linking out of a publishing house
- Linking into a publishing house
- Linking to deeper knowledge
- Linking infrastructure
- Linking all around

Elsevier's Linked Data Repository adds capabilities to share knowledge



Shouldn't our machines have access too?



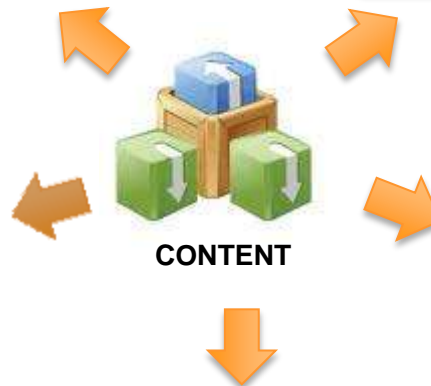
Run extensive searches and use locally loaded content for text mining purposes for their own research.

Perform extensive mining operations on subscribed content .

- Structuring input text
- Deriving patterns within the structured text
- Evaluation and interpretation of the output.



Extract semantic entities from Elsevier content for the purpose of recognition and classification of the relations between them



Enabling developers who wish to design and implement applications to analyse our content, or test applications as part of their research within Elsevier content



Integrate results on a server used for the customer's own mining system for access and use by its researchers through the customer's internal secure network.



Elsevier's New Text and Data Mining Policy (TDM)



Researchers at academic institutions can text mine subscribed content on ScienceDirect for non-commercial purposes via the ScienceDirect APIs

Access is granted to faculty, researchers, staff and students at the subscribing institution

Text mining output can be shared publically under these conditions

1. May contain "snippets" of up to 200 characters of the original text
2. Should be licensed as CC-BY-NC
3. Should include DOI link to original content

Corporate and other subscribers

- Your Elsevier Account Manager will be happy to discuss options to meet your needs

Open access content

- Text and Data mining permission are determined by the author's choice of user license. This information is detailed in the individual articles

Policy Aligned with the Recent STM Declaration on Text and Data mining



stm

INTERNATIONAL ASSOCIATION OF SCIENTIFIC, TECHNICAL & MEDICAL PUBLISHERS

www.stm-assoc.org

TEXT AND DATA MINING FOR NON-COMMERCIAL SCIENTIFIC RESEARCH

A STATEMENT OF COMMITMENT BY STM PUBLISHERS TO A ROADMAP TO ENABLE TEXT AND DATA MINING (TDM) FOR NON COMMERCIAL SCIENTIFIC RESEARCH IN THE EUROPEAN UNION

Recalling that the International Association of STM Publishers (STM) and member publishers have actively and constructively supported the discussions in the Licences for Europe Working Group Four,

Recognising that a high level of copyright and database protection, together with interoperability standards, technology innovations and sustainable business models, are vital for the viability of the creative industries.

Reaffirming that licensing is the smart and speedy route to providing access and the rights needed for text and data mining, and for related technologies such as text-to-speech and automated translation services

Supporting TDM at Elsevier



2006

- Began to support *ad-hoc* TDM access requests from customers

...

2012

- First Content Mining policy published
- New APIs and Content Syndication Service rolled out to provide better technical solutions for TDM content access

2013

- Pilot with ~30 academic customers to better understand needs and define future policy

2014

- New Text and Data Mining policy for academic customers announced

TDM Pilot Learnings – Use Cases



Most academic Mining requests fall in one or both of these categories:

1. Answering a specific research question

- How long does it take for concepts in STM literature to reach general media?
- What is the relationship between the research and consulting commitments of economics and finance professors?
- What are the characteristics of subjects in social psychology experiments?

2. Building a new data resource for the community

- An HIV mutation database for which mutations found in literature are mapped to the underlying database sequence
- A database on growth and alimentation of fishes, and develop a fish classification to identify new species for aquaculture
- A database with the electrophysiological properties of diverse neuron types

TDM Pilot Learnings – Researcher Challenges



Technical

- Obtaining necessary infrastructure
- Having to deal with different formats from content providers
- Sourcing and understanding TDM technology

Functional

- Fine-tuning pipeline, curating output, representing output meaningfully

Logistical/Legal

- Gaining access to the needed content
- Gaining permission to mine the content

TDM Pilot Learnings – Library Challenges



Expertise

- Understanding specific TDM-based projects well enough to assess implications & offer advice to library patrons

Legal

- Understanding and tracking what is allowed for what resources
- Negotiating permissions with multiple providers
- Ensuring academic freedom is protected

Financial

- Concerns about any additional costs
- Understanding how TDM affects usage figures for the library

Conclusions of the TDM Pilot:



1. Elsevier can offer enhanced value to ScienceDirect customers by including basic text and data mining access rights in subscription agreements
2. Self-service access to content for TDM via our APIs meets the needs of most researchers in academia
3. There is demand for services beyond basic content access that make text mining easier

What do Institutions need to do to get TDM Access NOW?



- TDM access clause will be part of standard ScienceDirect subscription agreement for new academic customers and upon renewal
- For existing agreements, an add-on contract amendment is available – ask librarian to contact your Elsevier Account Manager
- After signing institutional agreement/amendment, access to our API key registration page for your researchers will be enabled for your institution's IP address range

TDM Access: What do Researchers Need to do?



Use API Key to retrieve full text of journal articles and book chapters via the Elsevier API

- Elsevier XML and plain-text formats supported
- We are looking into supporting specialized text-mining friendly XML formats

Process retrieved full-text through text mining tools/workflow of choice

But Elsevier publishes only part of the scientific literature – and i need to mine all of it.



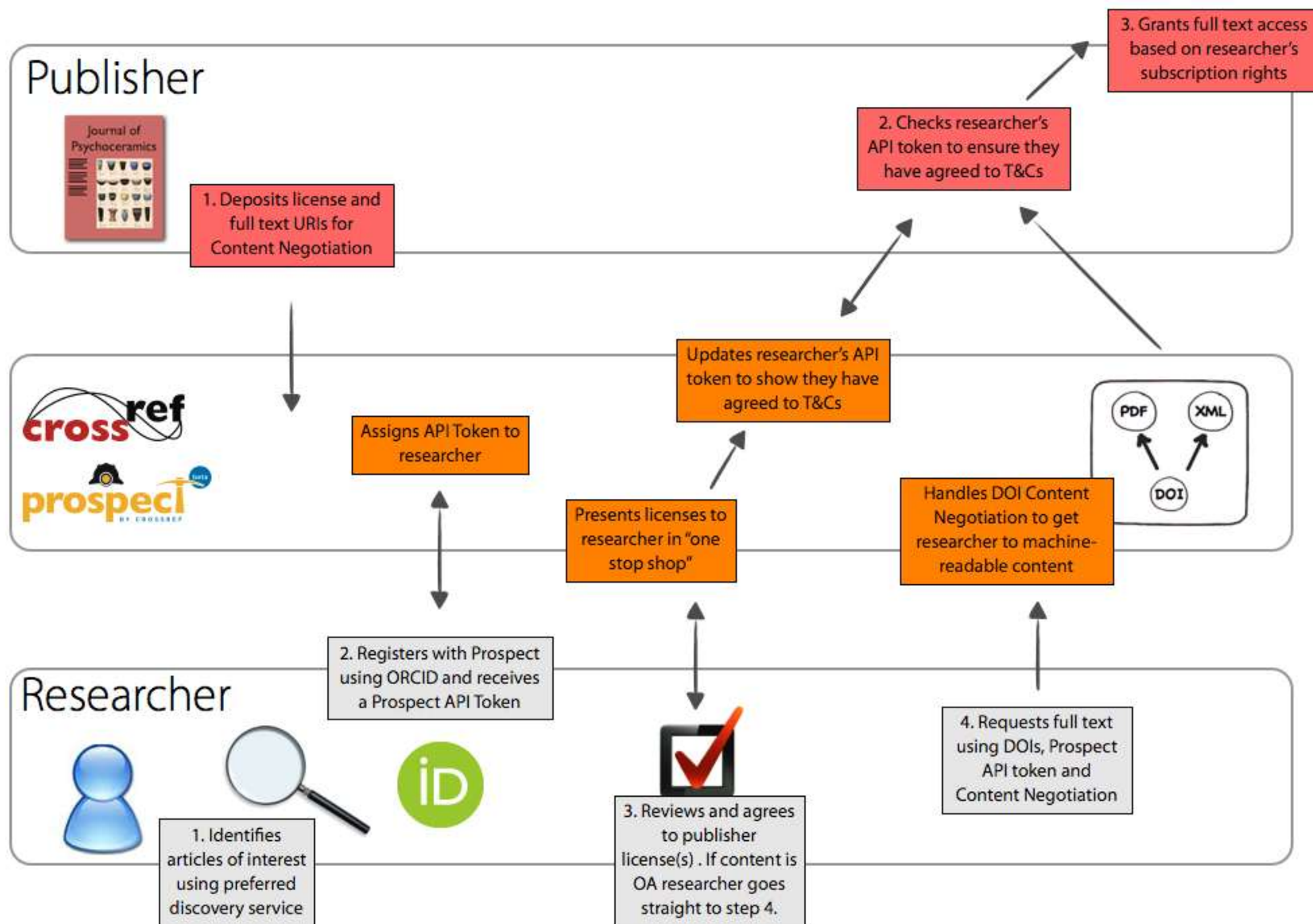
Supporting Cross-publisher TDM:



Prospect is a new service from CrossRef providing two components to address the issue of text and data mining scholarly literature across multiple publishers:

- The “Prospect Common API” (PCAPI) can be used to access the full text of content identified by CrossRef DOIs across publisher sites and regardless of their business model.
- The “Prospect License Registry” (PLR) can (optionally) be used by researchers and publishers as an efficient mechanism to provide “click-through” agreement of proprietary TDM licenses.
- Both components are free to use by researchers and the public

Prospect Workflow

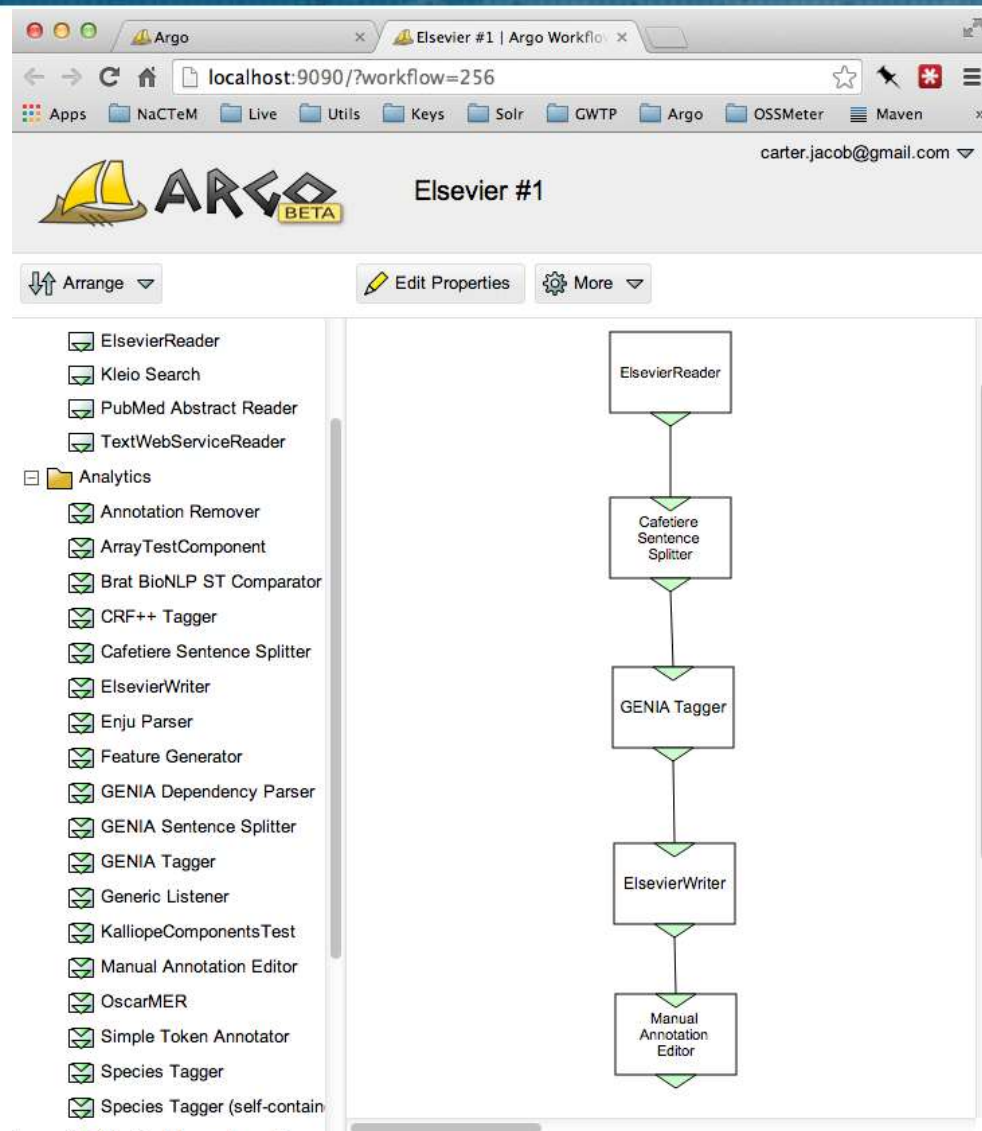


Elsevier is the first publisher to fully integrate with the Prospect beta system:

- Researchers may read and agree to the Elsevier TDM click-through agreement via the Prospect License Registry
- Researchers may use a Prospect token to access Elsevier content through the Prospect Common API rather than using the Elsevier-specific API Key and Elsevier API
- Content is available in the same formats as the Elsevier API

Text Mining as a Service

- Pilot with NaCTeM to integrate their tools with Elsevier content
- Hosted in the cloud
- Avoids the need for researchers to build and maintain TDM infrastructure
- Ability to define and execute TDM workflows in a graphical environment



Takeaway Points



- Researchers at academic institutions can now text-mine subscribed Elsevier content for non-commercial purposes at no additional cost
- Contact your librarian -> Elsevier Account Manager if you are interested
- Elsevier is collaborating with customers and industry partners to make text mining easier

Much may change, though some truths are likely to remain constant...



‘What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.’

– Herbert Simon, 1971

Much may change, though some truths are likely to remain constant...



Elsevier remains committed to publishing the highest quality research; now and into any future



<http://www.journals.elsevier.com/big-data-research/>

Thank you for your attention!



Further reading:

- Research Data Services:
<http://researchdata.elsevier.com>
- Database linking:
<http://www.elsevier.com/databaselinking>
- Article of the Future and Content Innovation:
<http://www.elsevier.com/about/content-innovation>
- Elsevier Developers portal (API for TDM):
<http://www.developers.elsevier.com/cms/index>
- The Executable Paper Pilot:
<http://www.elsevier.com/executablepaper>
- Semantic Web Challenge:
<http://challenge.semanticweb.org/>
- Big Data Research:
<http://www.journals.elsevier.com/big-data-research/>